STAT 651 Chapter 4

Zepu Zhang

December 16, 2010

1 Definition of multivariate random variables (or random vectors)

Definition 4.1.1, multivariate random variables.

Compare with definition 1.4.1, p. 27.

The multiple variates are either put in a vector (apparently the order matters), or individually named to distinguish.

2 Joint pmf and pdf

On page 140 appears again the definition of a discrete variable: it only has a countable number of possible values.

Definition 4.1.3, page 140. (Compare with definition 1.6.1, p. 34.)

Definition 4.1.10, page 144. (Compare with definition 1.6.3, p. 35.)

(We often use "dimension" to refer to components in multivariate distributions, e.g. "high dimensional density", "high dimensional random variable", "sum over this dimension", "integrate over/out that dimension".)

Generalization (4.6.1), (4.6.2), p. 177.

cdf p. 147, above section 4.2. Concept of cdf and its relation with pdf.

CDF for multivariate is not very useful, but you still need to know the concept. Generalization to more than two dimensions is straightforward.

3 Expectations of multivariate random variables

Totally analogous to the univariate counterpart.

Discrete:

$$Eg(X,Y) = \sum g(x,y)f(x,y)$$

Continuous:

$$Eg(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

A usual difficulty in the multivariate integral is that one variate affects the limits of integration over another variate.

Example 4.1.12, page 146.

Example Exercise 4.5(a), page 192.

Example Exercise 4.12, page 193.

On the one hand, the expectation of a n-dimensional random variable is a n-vector. On the other hand, we may define function $g(x,y)$ whose value is a single number, or a vector. Then the expectation of $g(X,Y)$ is calculated as above, which has the dimension of the value of $g(X,Y)$.

"Variance" of a multivariate variable is more involved than its univariate counterpart, and will be dealt with a little later. Basically, it will be "covariance".

Generalization (4.6.3), p. 178.

4 Marginal pmf and pdf

Theorem 4.1.6, page 143. Discrete case.

To find $P(X = x)$, we identify all possible values of the random vector where the X component is x (and accept whatever values for the other components), and sum up their probabilities.

$$P(X = x) = \sum_{y} P(x,y)$$

Theorem (4.1.3), page 145. Continuous case.

To find $f_X(x)$, we identify the line (or plane, if the joint has more than two components) with $X = x$ (and accept whatever values for the other dimensions), and integrate the joint pdf along this line (or plane). This is called integrate out the other variates, or marginalize.

Note that the above are not "definitions" of marginal pmf and pdf (but can almost be taken as such). The definition, informally, is the pmf (or pdf) of X, simply ignoring the other variates.

Note 1. The marginal pmf (pdf) is a pmf (pdf), hence has all the necessary properties of a pmf (pdf), mainly (1) non-negative; (2) sum (integrate) to 1.

2. If we need to discuss a variate regardless of the joint ones, we need to get its marginal. Then this is just a univariate variable, and we can calculate probabilities, expectations, summaries, etc.

3. The joint distribution contains full information; especially, we can derive the marginals from the joint, if needed. But we can not derive the joint from the marginals. Indeed, given the marginal for each variate, the joint does not need to be unique (and this means the same as that we can not determine the joint). While going from joint to marginal, we sum over (or integrate out) all other variates, and that information can not be recovered.

4. Understand continuous joint density geometrically via "volume", just like understanding univariate density via "area".

Example 4.1.7, page 143.

Example 4.1.11, page 145.

Generalization (4.6.4), (4.6.5), p. 178.

If the "full" rv has more than two dimensions, then a "marginal" itself can be multivariate, e.g. from the joint $f(x, y, z)$, we may get the marginal $f(x, y)$ by integrating over z. This marginal is really the "joint" of X and Y.

5 Conditional pmf/pdf

Definition 4.2.1, page 148.

Definition 4.2.3, page 150.

Note 1. The basic pattern is

$$\text{conditional} = \frac{\text{joint}}{\text{marginal}}$$

where the "marginal" is that of the conditioning variable. The probability analogy is

$$P(A \,|\, B) = \frac{P(A \cap B)}{P(B)}$$

2. Note the condition: the conditional is defined only for a conditioning value where the marginal is > 0.

3. Again, the conditional, for a fixed conditioning value, is a pmf/pdf, therefore it has all properties of a pmf/pdf, and can be used to calculate probabilities, expectations, all kinds of summaries, etc. (1st paragraph, page 151.)

4. In the discrete case, say $P(Y \mid X = x)$, the pmf is the joint probabilities $P(x, y)$, where X is fixed at x, normalized so that the probabilities sum to 1. The normalizing constant is the marginal $P(X = x)$. Imagine you identify all possible values (x, y) with fixed x and various y, keep their relative probabilities and scale them so that they sum to 1.

5. In the continuous case, say $f(y \mid X = x)$, the pdf is the joint density curve $f(x, y)$, where X is fixed at x, normalized so that the area under the curve integrates to 1. The normalizing constant is the marginal $f_X(x)$. Imagine you transect the joint pdf surface at $X = x$, keep the shape of the profile curve but divide it by its area below so that it is a pdf.

6. Compare the concepts of marginal and conditional: marginalization involves summing/integrating over other variates, whereas conditioning involves cutting and normalizing. Marginalization does not need normalization—the sum or integral will turn out to be a valid pmf/pdf. Conditioning does not do summation or integration—it simply takes a transect of the joint. Marginalization is a “sum” operation because the other variates can be any value and we need to combine their probabilities. Conditioning is a “profiling” operation because the conditioning dimension is now fixed.

Example 4.2.2, page 148, calculating conditional pmf.

Example 4.2.4, page 150, calculating conditional pdf.

One way to define/specify/construct a joint pdf is: define the marginal $f_X(x)$ and conditional $f(y \mid x)$ for all x values, then the joint is $f(x, y) = f_X(x) f(y \mid x)$.

Generalization (4.6.6), p. 178.

The condition part and the conditional (or conditioned) part can be univariate or multivariate, as (4.6.6) shows.

6 Independence

If the conditional is the same as the marginal, i.e. the condition contribution no information, then X and Y are independent.

Definition 4.2.5, page 152.

$$f(x, y) = f_X(x)\, f_Y(y)$$

Note 1. Note the condition: the above must hold for all values of X, Y.

2. However, in the continuous case this relation is allowed to fail at a countable number of values; see the paragraph on page 156 below theorem 4.2.14. Recall that "spikes" on the pdf curve does not affect probabilities. Do not worry too much about this point.

2. Independence ⇔

$$\text{joint} = \prod \text{marginals}$$

3. To check independence,

1. In the discrete case, if the pmf is given via a complete list of all joint probabilities, then verify the defining relation for all values of X and Y. You need to derive the marginals for this purpose. Example 4.2.6, page 152.

2. If the joint is given by a formula (continuous distributions is always always given this way), use theorem 4.2.7, p. 153. Example 4.2.8, page 153.

Example Joint normal pdf.

3. Use theorem 4.3.5, p. 161.

4. Not independent if the support of the joint is not a rectangular box; see second paragraph on page 154. Such "cross-product" support is a necessary but not sufficient condition for independence. The support is not "rectangular" if the joint pmf/pdf depends on a relation between the variates, e.g. $X > Y$.

4. If, based on subject-matter knowledge, we know two variables are independent, then we can specify the joint pmf/pdf using the product of the marginals. Example 4.2.9, page 154.

Caution: pairwise independence is not enough for mutual independence of multiple variables!

Theorem 4.2.10, p. 154.

Note 1. Very useful! Part a is general and useful. Part b is more specialized and its application arises a lot.

2. Both results stem from the decomposition of the joint pmf/pdf into product of marginals, so that summation/integration

is a product of summations/integrals with no interferences between the variates.

Example 4.2.11, p. 155.

Theorem 4.2.12, p. 155.

Theorem 4.2.14, p. 156.

Note 1. This is an important and useful result, but we'll make it more general later. There is not much need to memorize this particular theorem about the sum of exactly two independent normal variables.

2. A general statement that we will learn later says a linear combination of normal variables is normal. Independence is NOT required.

3. In this theorem, the mean $E(X+Y) = E(X) + E(Y)$ is always true regardless of the distributions and independence. The variance $\text{var}(X+Y) = \text{var}(X) + \text{var}(Y)$ is a result of independence, otherwise it hold for any distributions. The really special element of this theorem is that the resultant variable (the sum) is still normal.

7 Mutual independence

Definition 4.6.5, p. 182. The "mutual independence" between more than two rv.

This generalizes definition 4.2.5, p. 152.

Note 1. Following the definition, any subset of the variables are mutually independent. This can be seen by getting the marginal of the subset, which again is in a form the conforms to the definition above.

2. In particular, pairwise independence follows from mutual independence.

3. However, the other direction is not correct. Pairwise independence does not guarantee mutual independence of a larger group of rv's.

Generalization Theorem 4.6.6, p. 183, generalizes 4.2.10, p. 154.
Theorem 4.6.7, p. 183, generalizes 4.2.12, p. 155.
Theorem 4.6.11, p. 184, generalizes 4.2.7, p. 153.
Theorem 4.6.12, p. 184, generalizes 4.3.5, p. 161.

Generalization Corollary 4.6.9, p. 183, generalizes theorem 4.2.12, p. 155. But this result can be derived on the fly, so don't try to memorize it.

Generalization Corollary 4.6.10, p. 184, generalizes theorem 4.2.14, p. 156. But like the latter, we will learn more general and elegant results about the normal distribution.

8 Distribution of functions of random variables, i.e. distribution of a transform

Section 4.3 should be compared to section 2.1.

8.1 Discrete case

Suppose X and Y are discrete r.v. with a known joint pmf. Now we have two other discrete r.v defined as functions of X and Y:

$$U = g(X, Y), \quad V = g(X, Y)$$

How can we get the pmf of (U, V)?

The idea is entirely analogous to the univariate case on page 48. Basically, for $(U, V) = (u, v)$, we need to identify what (X, Y) values will result in this (U, V) value, and the total probability of these (X, Y) values is the probability for (U, V) to assume the value (u, v).

See (4.3.1), page 157.

Example X and Y are independent Poisson variables with parameters λ and θ, respectively. What is the distribution of $U = X+Y$?

First off, U is discrete and its possible values are $0, 1, \dots$ Then we only need to find the probability that U takes any of these values.

Notice that for U to be u, X can be any value from 0 up to

u, then Y must be correspondingly $u - x$.

$$\begin{aligned}
P(U = u) &= \sum_{x=0}^{u} P(X = x)P(Y = u - x \mid X = x) \\
&= \sum_{x=0}^{u} P(X = x)P(Y = u - x) \\
&= \sum_{x=0}^{u} \frac{e^{-\lambda}\lambda^x}{x!} \frac{e^{-\theta}\theta^{u-x}}{(u-x)!} \\
&= \frac{e^{-\lambda-\theta}}{u!} \sum_{x=0}^{u} \binom{u}{x} \lambda^x \theta^{u-x} \\
&= \frac{e^{-\lambda-\theta}}{u!} (\lambda + \theta)^u \\
&\sim \text{Poisson}(u \mid \lambda + \theta)
\end{aligned}$$

Also check out the alternative approach in example 4.3.1, p. 157.

Generalization Generalization to more than two dimensions is straightforward.

8.2 Continuous case

Now we have a formula analogous to theorem 2.1.5, p. 51.

Concept Jacobian of a transformation, page 158.

Theorem (4.3.2), page 158.

Note This formula requires the transformation to be "one-to-one" and "onto". If not, we'll work in subsets within which the transformation is so, then piece together the results. This is the generalization in (4.3.6) on page 161, analogous to theorem 2.1.8, p. 53.

Example 4.3.4, p. 159.

Example 4.3.6, p. 162.

Generalization (4.6.7), p. 185.

Example 4.6.13, p. 185.

9 Multinomial distribution and multinomial theorem

Definition 4.6.2, p. 180. Multinomial distribution.

This generalizes binomial, top of page 90. The kind of application problems are also analogous to binomial.

Theorem 4.6.4, p. 181.

This generalizes the binomial theorem, p. 90.

Proof: use induction.

Example 4.6.3, p. 181.

10 Covariance and correlation

10.1 Concepts

Definition 4.5.1, p. 169.

Definition 4.5.2, p. 169.

Note 1. The value of cov depends on the unit. It is unbounded, hence its value does not tell us how strong the relationship is.

2. The sign of cov tells us whether the two variables are positively or negatively related.

3. Corr is cov normalized by standard deviation. Its value does not dependent on the unit, and it is bounded on $[-1, 1]$. The corr tells us the strength (in addition to the direction) of the relationship.

4. Cov and corr reflect linear relationship.

10.2 Properties

Theorem 4.5.3, p. 170. Important!

Cf. (2.3.1), p. 60.

Theorem 4.5.5, p. 171. Important!

Note 1. Independence ⇒ zero cov; but the converse is not true.

2. If X and Y are jointly normal, then the converse is true.

Theorem 4.5.6, p. 171. Important!

Generalization

$$\mathrm{var}(a_1 X_1 + \cdots + a_n X_n) = \sum_{i,j} a_i a_j \,\mathrm{cov}(X_i, X_j)$$

Note $\mathrm{cov}(X, X) = \mathrm{var}(X)$. This handles the cases where

$i = j$. If the X_i's are independent, then

$$\text{var}(a_1X_1 + \cdots + a_nX_n) = \sum_i a_i^2 \,\text{var}(X_i)$$

How can you prove these generalizations? (Hint: induction.)

Theorem 4.5.7, p. 172.

Example 4.5.4, p. 170.

Example 4.5.8, p. 173.

Example 4.5.9, p. 174.

10.3 Generalizations

Covariance describes the relationship between two random variables. It is a fundamental concept in statistics. It is also different from the other concepts (joint pdf, marginal, conditional, independence, etc.) in how it generalizes to more than two variables. We do not have something like "covariance between three variables". We only have covariance between two variables. For three variables, we define their covariance matrix as composed of pair-wise covariances.

$$\text{cov}(X, Y, Z) \equiv \begin{bmatrix} \text{cov}(X,X) & \text{cov}(X,Y) & \text{cov}(X,Z) \\ \text{cov}(Y,X) & \text{cov}(Y,Y) & \text{cov}(Y,Z) \\ \text{cov}(Z,X) & \text{cov}(Z,Y) & \text{cov}(Z,Z) \end{bmatrix}$$

Let $\vec{X}$ be a n-vector, then

$$\text{cov}(\vec{X}) \equiv [\text{cov}(X_i, X_j)]_{i,j} = E\Big\{\big(\vec{X} - E(\vec{X})\big)\big(\vec{X} - E(\vec{X})\big)^T\Big\}$$

The ith diagonal element is $\text{var}(X_i)$. If X_i and X_j are independent of each other, then the (i, j) element is 0. If all the X's are mutually independent, then only the diagonal elements are nonzero.

Also,

$$\text{cov}(\vec{X}) \equiv [\text{cov}(X_i, X_j)]_{i,j} = \big[E(X_iX_j) - E(X_i)E(X_j)\big]_{i,j} = E(\vec{X}\vec{X}^T) - (E\vec{X})(E\vec{X})^T$$

The inverse of a cov matrix is important. Happens a lot in analysis and applications.

The determinant of a cov matrix is also important.

The most important property of a cov matrix is that it is positive definite.

11 Multivariate normal distribution

Definition 4.5.10, p. 175, and the rest: you only need to known such a formula for bivariate normal exists and can be found. Don't bother to memorize it. I've never used this formula. Instead, we'll use the general concepts below.

Let $\vec{X}$ be a random vector of p dimensions. Suppose $\mathbf{\Sigma}$ is the cov matrix of $\vec{X}$ and $\vec{\mu}$ is the mean vector of $\vec{X}$. The random vector $\vec{X}$ has a multivariate normal distribution if its density is the following:

$$f(\vec{x}) = \frac{1}{\sqrt{(2\pi)^p|\mathbf{\Sigma}|}} \exp\Big\{-\frac{1}{2}(\vec{x}-\vec{\mu})^T\mathbf{\Sigma}^{-1}(\vec{x}-\vec{\mu})\Big\}$$

Note the univariate normal density is a special case of this.

Properties 1. **Marginal**. All marginals are normal.

2. **Linear transform**. Let $\boldsymbol{A}$ be a $n \times p$ matrix, where $n \leq p$, then

$$\boldsymbol{A}\vec{X} \sim N\big(\boldsymbol{A}\vec{\mu}, \boldsymbol{A}\mathbf{\Sigma}\boldsymbol{A}^T\big)$$

2. **Conditional**. Let $\vec{X} = \begin{bmatrix}\vec{X}_1\\ \vec{X}_2\end{bmatrix}$ be distributed as $N(\vec{\mu}, \mathbf{\Sigma})$ with $\vec{\mu} = \begin{bmatrix}\vec{\mu}_1\\ \vec{\mu}_2\end{bmatrix}$ and $\mathbf{\Sigma} = \begin{bmatrix}\mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12}\\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22}\end{bmatrix}$, and $|\mathbf{\Sigma}_{22}| > 0$. Then given $\vec{X}_2 = \vec{x}_2$, the conditional distribution of $\vec{X}_1$ is

$$\vec{X}_1 \,|\, \vec{x}_2 \sim N\big(\vec{\mu}_1 + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}(\vec{x}_2 - \vec{\mu}_2),\ \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}\big).$$

12 Hierarchical models and mixture distributions

Example 4.4.1, 4.4.2, p. 163.

The advantages of hierarchical models: a complex problem is divided into hierarchies; each hierarchy is relatively simple in terms of understanding (it is relatively easy to come up with a model for it; it is relatively easy and intuitive to judge whether the model is reasonable, often because the model describes mechanisms in the actual problem, that is, the model has understandable interpretation) and mathematics.

Definition 4.4.4, p. 165. Mixture distribution

Almost always, a distribution has parameters. The distribution is determined if the parameters have fixed values. If the parameters are themselves random with a certain distribution, then we have a hierarchy: from the parameter's distribution comes a specific value of the parameter; with this specific value of parameter, the final target distribution is fixed.

This target distribution is said to be a mixture. Therefore a hierarchical model leads to a mixture distribution.

Example A binomial model relies on the rate parameter p. Suppose this p is uncertain, and we assume it has a beta distribution.

Example Suppose a random number is from $N(0, 2)$ with probability 0.3 and from $N(1, 3)$ with probability 0.7. Then its density is

$$f(x) = 0.3\frac{1}{\sqrt{2\pi \cdot 2}}e^{-x^2/4} + 0.7\frac{1}{\sqrt{2\pi \cdot 3}}e^{-(x-1)^2/6}$$

To fit in the definition 4.4.4, we can say the distribution of X depends on a random quantity Y, which is 0 with probability 0.3 and 1 with probability 0.7. If $y = 0$, then the distribution of X is $N(0, 2)$; if $y = 1$, then the distribution of X is $N(1, 3)$. This hierarchy helps understanding, for example it may correspond to the fact in the particular application that X comes from two sources with relative frequencies 0.3 and 0.7. The distribution of X in the end is not uncertainty; it is clearly defined to have the density function above.

Another way to understand it—X has a normal distribution with parameter vector $\theta = (\mu, \sigma^2)$ but the value of θ is not fixed. It has two possible values, $(0, 2)$ and $(1, 3)$, with probabilities 0.3 and 0.7, respectively.

This kind of normal mixture is very flexible and versatile. It has important applications.

Theorem 4.4.3, p. 164.

$$\text{mean} = \text{mean of conditional mean}$$

Read the comments lower on page 164.

Theorem 4.4.7, p. 167.

variance = variance of conditional mean + mean of conditional variance

Note both terms on the right-hand side are non-negative.

There is a similar-looking relation that is very commonly used in some studies. Suppose something has a fixed unknown value a. We model (or estimate) a by a random variable X which is <u>biased</u>. A standard measure of the quality of X is the “mean squared error”, that is, $E(X - a)^2$. We have

$$\begin{aligned}
E(X - a)^2 &= E(X - \mu_X + \mu_X - a)^2 \\
&= E(X - \mu_X)^2 + 2(\mu_X - a)E(X - \mu_X) + (\mu_X - a)^2 \\
&= \text{var}(X) + (\mu_X - a)^2
\end{aligned}$$

Note $\mu_X - a$ is the bias of X in estimating a. This relation is the "variance-biased decomposition" of mean squared error:

$$\text{mean squared error} = \text{variance} + \text{bias}^2$$

This is a side track. It's something useful to know.

(Typo: p. 167, line 3, (3.2.10) should be (3.3.10).)

13 Inequalities

It is hard to have an intuitive grasp of the inequality relations listed in this section without having used them in work. For now you only need to know the following.

Theorem 4.7.3, Cauchy-Schwarz inequality:

$$|EXY| \le E|XY| \le \sqrt{(EX^2)(EY^2)}$$

The first inequality comes from the linear properties of the expectation (theorem 2.2.5, p. 57), noticing $XY \le |XY|$. The second inequality is a special case of Hölder's inequality.

Example 4.7.4, covariance inequality:

$$\big(\operatorname{cov}(X,Y)\big)^2 \le \operatorname{var}(X)\operatorname{var}(Y)$$

Note this is an application of Cauchy-Schwarz: take X to be $X - \mu_X$ and Y to be $Y - \mu_Y$ in Cauchy-Schwarz.

With this relation, it is straightforward to prove that the correlation coefficient, defined as $\rho(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}}$ is bounded by -1 and 1.

Example (4.7.5), p. 188. A special case of Hölder's inequality is

$$E|X| \le \big(E(|X|^p)\big)^{1/p}, \quad p > 1$$

Taking $p = 2$, we get

$$E|X| \le \sqrt{EX^2}$$

that is,

$$E(X^2) \ge (EX)^2$$

Recall $\operatorname{var}(X) = E(X^2) - (EX)^2 \ge 0$, consistent with this result.

The relation in the middle of page 189, before section 4.7.2, is similar. It states that for numbers a_i, $i = 1, \dots, n$,

$$\frac{1}{n}\Big(\sum a_i\Big)^2 \le \sum a_i^2$$

i.e.

$$\left(\frac{\sum a_i}{n}\right)^2 \le \frac{\sum a_i^2}{n}, \quad \text{or } (\overline{a})^2 \le \overline{a^2}$$

Definition 4.7.6, p. 189. Convex and concave functions.

Understand this concept in several ways:

1. Have a picture of the curve: a bowl is convex, a bell is concave, x^2 is convect, $\log x$ is concave, etc.

   (Remember the concept by this picture; the others below can be understood through this picture.)

2. Tangent lines are below the curve of a convex function.

3. Connecting two points on a convex function, the line is above the function's curve.

4. If the second derivative exists, then for a convex function $g(x)$, the second derivative is ≥ 0 for all x. On a convex curve, "glide" a tangent line as x increases, we see the slope of the tangent line increases, hence the second derivative is positive (or at least nonnegative). Recall the second derivative reflects the change of the first derivative.

   (This is the way to go if we are to check whether a function is convex or concave, given an analytical form of the function.)

Theorem 4.7.7, Jensen's inequality, p. 190.

$$\begin{aligned} Eg(X) &\ge g(EX) \text{ if } g(x) \text{ is convex} \\ Eg(X) &\le g(EX) \text{ if } g(x) \text{ is concave} \end{aligned}$$

Proof.

Understand the condition for "=" to hold—It basically means for "=" to hold, the function $g(x)$ is linear. As long as it curves, the inequality is strict. (Of course, the statement requires $g(x)$ to be convex or concave, not just curvy.) More accurately, the condition says that the function leaves a straight line with probability 0. In other words, if some points of the function are off the straight line, they must be just discrete points so that they have 0 probability mass. (The condition in theorem 4.5.7, p. 172, is similar.)

Example: $EX^2 \ge (EX)^2$; $E(1/X) \ge 1/EX$ (where $X > 0$).

Example 4.7.8, p. 191.

harmonic mean $\le$ geometric mean $\le$ arithmetic mean